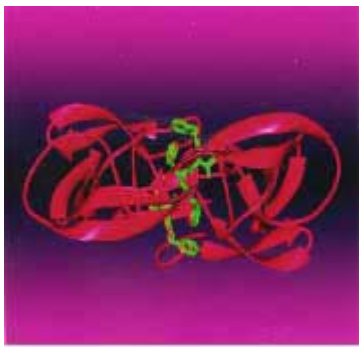
Rapid Virtual Screening of Large Drug Databases

A Research Project of Dr. Jack R. Collins and Dr. Brian T. Luke

In the last 2 years the number of structures in the Brookhaven protein data bank has jumped from 3964 to 6828. The detailed knowledge of the atomic structure of enzymes and proteins has fueled the investigation of new potential drug targets by the application rational drug design strategies. This combination of newly determined protein structures, combinatorial chemistry, and high throughput screening has led to an explosion in the number of possible therapeutic targets and potential drug leads. Often, the investigation of new targets is hampered by the absence of initial lead compounds that bind to a specific site in the protein, have reasonable affinity (typically < 5 micromolar), and can be easily modified to generate analogues for subsequent testing and optimization. High throughput screening (HTS) has been applied in many cases where an assay suitable for HTS is known. For farnesyl protein transferase, researchers at Schering-Plough have shown, however, that computer-based 3D screening methods can yield a larger percentage of "hits" than HTS at a fraction of the cost. In cases where the structure of a protein is known, the computer screens of molecular 3D databases, such as the NCI or Available Chemicals Directory (ACD) databases, can suggest initial lead molecules that are readily available. It is the goal of this research to develop a series of computer programs that can rapidly and reliably scan a 3D database to suggest the most likely candidates for laboratory testing.



F-1. Docking of A76928 into the HIV-1 protease cavity. The experimentally determined position is shown in green and the docked structure in gold.

Currently available programs used for these types of problems (e.g. DOCK, AUTODOCK, HAMMERHEAD, FLEXX) are either slow, impose severe limitations on the types of ligands that can be considered, or are proprietary. The goal of our development effort is to create a parallel 3D-docking and scoring algorithm that can be integrated into a suite of programs to efficiently screen large structural databases against protein targets for lead generation and subsequent testing. To accomplish this we have chosen a grid-based linear response function coupled with an evolutionary searching algorithm that scales linearly with this number of atoms in the molecule being docked. We emphasize that our goal is to develop a rapid rather than an exhaustive search.

Success

F-2. An overlap of the lowest energy nevirapine docked structures (red) compared to the crystallographic position. The docked structures are displaced approximately 0.4A RMS, a feature primarily due to the exclusion of 2 bound water molecules forming bridges between nevirapine and the protein.

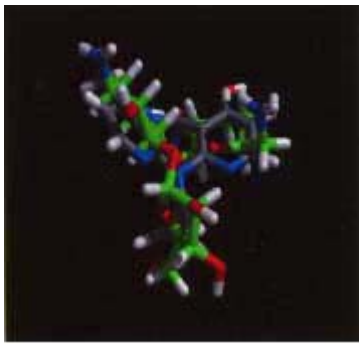
The features and use of our docking and database search programs are illustrated in two examples. In the first we compared the results of our program with the known crystal structures of 10 HIV-1 protease inhibitors complexed with the enzyme, as well as nevirapine in its HIV-1 reverse transcriptase (RT) binding site. In the second example, we searched the NCI 3D database to find molecules that fit well into the RT-nonnucleoside inhibitor (NNI) binding site and could serve as possible lead compounds in a drug development effort.

One of the most intensely studied targets of computer-aided, structure-based, drug design efforts in the past several years has been HIV-1 protease. As a result, the crystal structures of several drug candidates have been solved and published. We chose ten of these molecules as test cases for our rigid docking procedures. Based on the crystallographic structures, our docking procedures resulted in an average RMS deviation of 0.39Å from experiment. These results indicate that it is possible to obtain computer docked structures that are virtually identical to experiment when the binding conformation is known.

We are currently working on flexible docking procedures while trying to maintain the goal of a rapid screen. Programs such as DOCK 4.0 will not perform flexible docking on the ten molecules used in our study since they do not have proper anchor points and have more than eight rotatable bands. The current algorithms we employ dock each rigid structure in approximately 2.5 seconds when run on a DEC-alpha 440MHz processor. With further algorithmic development and increases in processor speed we hope to be able to perform several dockings per second. Increased speed is necessary to realistically accomplish flexible docking of drug-like molecules. This need for speed is even more evident in the next example where we screen the NCI database for compounds that fit the HIV-1 RT NNI binding site.

We started from the nevirapine bound HIV-1 RT crystal structure and first determined whether our program would correctly dock the inhibitor back into the protein. The results of our docking procedure are shown Figured F-2.

Based on the success of the nevirapine results, we performed a screen of the NCI 3D database in the same binding cavity as nevirapine. Approximately 75% of the compounds in the database passed our pre-screen and were actually docked into the RT binding site. This took almost eight seconds per compound with the program running on an IBM SP-2 containing 66MHz Power2 chips. Based on the static structures contained in the NCI database, we found several compounds that gave good "fits" in the RT binding cavity. One compound that scored particularly well is vistamycin (shown in Figure F-3).



F-3. The docked orientation of vistamycin (shown with green carbon atoms) along with the experimental positional of nevirapine (shown with gray carbons). The geometric and shape similarities are quite striking.

Benefits of Scalable Increases in Compute Power

The utility of molecular docking programs has been severely limited by the time it takes to perform a full database screen. This is illustrated by the timings taken from the current version of our program (which is one of the fastest available). The table below shows the computer power necessary to perform routine searches on databases ranging from the NCI 3D database to those available in the near future from the American Chemical Society (ACS). Two different time estimates are given: the first is for static (single conformation) searches and the second is for flexible compounds in rigid cavities. It is clear that the most useful results would be from the flexible dockings. The estimates are based on the speed of a single 195MHz SGI R10k CPU, and with a 64 processor parallel version shown in parentheses.

#Compounds	Static	Flexible
120,000	3 days	~150 days (~2 days)
1,000,000	25 days (~9 hrs.)	~1250 days (~20 days)
7,000,000	175 days (~3 days)	~8750 days (~140 days)

As the number of molecules in the molecular databases now maintained by vendors such as Beilstein and ACS increases, computer power will need to be increased dramatically to fully exploit the available information resources. In the meantime, further advances in algorithms for molecular docking and database searches will help to bring computer-based lead generation to reality.